LEVEL II 12

AD A101719

Computer-based Education Research Laboratory

15 DAHC 15-73-C-0077, ARPA Order-2245

University of Illinois Urbana Illinois

6 APPROACHES TO VALIDATION OF CRITERION-REFERENCED TESTS AND COMPUTER-BASED INSTRUCTION IN A MILITARY PROJECT

10 KIKUMI TATSUOKA

14 MTC ~~REPORT No.~~ 22

12 67

11 JANUARY 1978

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER MTC 22	2. GOVT ACCESSION NO. AD-A107 719	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Approaches to Validation of Criterion-Referenced Tests and Computer-Based Instruction in a Military Project		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) K. Tatsuoka		8. CONTRACT OR GRANT NUMBER(s) DAHC-15-73-C-0077
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Illinois Computer-based Education Research Laboratory		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS ARPA Order 2245
11. CONTROLLING OFFICE NAME AND ADDRESS DARPA 1400 Wilson Boulevard Arlington, Virginia 22209		12. REPORT DATE January 1978 13. NUMBER OF PAGES 65
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) NA		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document is approved for public release and sale; distribution is unlimited. This document may be reproduced for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Computers
Education
PLATO

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(U) The PLATO Air Force Base Computer-Based Education (PLATO AFB CBE) project at Chanute adopted the mastery learning technique in their 34 lessons and set the mastery criterion at 80% correct on the end of lesson test. They used the performance result of each criterion-referenced test (CRT) in two different ways: (1) for assessing the individual performance, and (2) for evaluation, or more precisely within Chanute's context, lesson evaluation.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(U)
The adoption of a criterion-referenced testing approach to evaluation raises two measurement issues that have relatively less importance in norm-referenced testing. The issues are: (1) definition of mastery, and (2) a priori standards. These issues still remain unsolved, but are receiving increasing attention.

(U) One purpose of this paper is to examine the appropriateness of the use of CRTs as a means of controlling an individual student's advancement to the next level of instruction or retainment in the current unit of instruction in the PLATO AFB CBE Program (~~or~~ project) at Chanute.

(U) Our other purpose in this paper is to turn the focus from the aspect of individual assessment to that of program evaluation, which requires the establishment of a criterion rate for validation of a lesson, so that a lesson would be considered validated if the percentage of failure rate at the end of the lesson was less than the criterion.

APPROACHES TO VALIDATION OF CRITERION-REFERENCED TESTS AND COMPUTER-BASED INSTRUCTION IN A MILITARY PROJECT

MTC Report No. 22

Kikumi Tatsuoka

January, 1978

COMPUTER-BASED EDUCATION RESEARCH LABORATORY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

This research was supported in part by the Advanced Research Projects Agency of the Department of Defense under U.S. Army Contract DAHC-15-73-C-0077.

ACKNOWLEDGMENTS

I wish to acknowledge the services rendered by the following persons:

Bob Baillie, expert programmer and numerical analyst, who has several publications in the Mathematics of Computation and other journals, wrote most of the main programs.

Tamar Weaver, a highly experienced former programmer at the Ministry of Transportation of Israel, who wrote several statistical analysis routines and transformation programs.

Kay Tatsuoka, junior in mathematics and computer science at the Massachusetts Institute of Technology, who wrote several statistical analysis routines and utility programs.

Jerry Dyer, Mark Bradley, and John Matheny, who served as junior programmers.

Julie Garrard, for editorial work and clerical help.

Curtis Tatsuoka, 15 years of age, for refining screen displays and carrying out other small programming tasks.

Professors Maurice Tatsuoka and Robert Linn, for their discussions and comments.

Roy Lipschutz and Wayne Wilson, for their artwork.

TABLE OF CONTENTS

CHAPTER

INTRODUCTION

The PLATO Air Force Base Computer-Based Education (PLATO AFB CBE) project at Chanute adopted the mastery learning technique in their 34 lessons and set the mastery criterion at 80% correct on the end of lesson test. They used the performance result of each criterion-referenced test (CRT) in two different ways: (1) for assessing the individual performance, and (2) for evaluation, or more precisely within Chanute's context, lesson evaluation.

The adoption of a criterion-referenced testing approach to evaluation raises two measurement issues that have relatively less importance in norm-referenced testing. The issues are (1) definition of mastery, and (2) a priori standards. These issues still remain unsolved, but are receiving increasing attention. A large number of articles relating to this subject have been published, but the many definitions of mastery are by no means equivalent. The concerns of these articles are limited to the use of criterion-referenced testing for individual assessment, i.e., judging whether or not a given student has mastered a given instruction to be learned to some suitable level of mastery (Block,1971; Emrick, 1971; Millman, 1973; Besel, 1971; Novick & Lewis, 1974; Roudabush,1974; Huynh, 1976; Linn,1977).

One purpose of this paper is to examine the appropriateness of the use of CRTs as a mean of controlling an individual student's advancement to the next level of instruction or retainment in the current unit of instruction in the PLATO AFB CBE Program (or project) at Chanute.

Our other purpose in this paper is to turn the focus from the aspect of individual assessment to that of program evaluation, which requires the establishment of a criterion rate for validation of a lesson, so that a lesson would be considered validated if the percentage of failure rate at the end of the lesson was less than the criterion.

Although there is a mathematical duality in both aspects of criterion-referenced testing, it is true that the program evaluation aspect has not received all the attention that it deserves. One reason for this is that the results of evaluation may call for expensive revisions in instructional materials, at least in traditional teaching settings. However, PLATO provides an ideal situation for program evaluation because revision of lessons can be done with relatively little trouble and expense. Therefore, it is important and necessary to explore reliable methods that will help to improve the quality of CAI lessons.

CRITERION-REFERENCED TEST AS ASSESSMENT OF PROGRAM EVALUATION

2.1 Mastery Learning Strategy

Mastery learning strategies have been used in many educational settings since Bloom (1968) advocated them in the late 1960's. In this new approach to instruction, a mastery level is set for the material to be learned so that a majority of the students must attain the criterion level.

Interesting findings about mastery learning strategies were reported by Carroll (1963), Atkinson (1968) Block (1970), Kim, Hogan, et al. (1970, 1971) and many others. According to Block (1971), mastery learning allowed 75-90% of students to achieve the same level as the top 25% of students in usually achieved with typical grouped instructional methods such as in regular class rooms.

A similar study by Kim et al. (1970, 1971) showed that 72% of approximately 5800 students in foreign language classes achieved a mastery criterion of 80% correct on final tests under the mastery condition while only 28% of the traditional condition achieved this level. The high percentage of students achieving criterion in the mastery condition shows the effectiveness of this strategy of instruction. However, these results may also be due partly to the quality of lessons given to the students during the experiment, or may even be due to the kinds of tests that were given to the students in order to examine the degree of mastery achieved in the instructional unit to be learned. We may be able to say that the high quality lessons produce a higher percentage of success than do low quality lessons if the tests given at the end of the lessons are comparable to one another.

The experienced instructional designer might say that the quality of instruction may be determined by the appropriateness of instructional cues and the quality and types of reinforcement given each student, as well as the amount of participation and practice experienced by each student. Therefore, determining the quality of instruction is a multidimensional and complicated task. It is very difficult to measure these factors and develop a method of setting validation criteria for CAI lessons based on the quantitative data from such complex variables. Since our concern is to restrict the discussion to the quantitative method of setting the validation criterion of a given lesson, we will start examining the validation criterion that has been used in the army, and the PLATO AFB CBE Program at Chanute Air Force Base.

2.2 Validation Criterion of Lessons in PLATO AFB CBE Program

The PLATO IV computer-based education system, in development for over a decade at the University of Illinois, was used in the training program of Special and General Purpose Vehicle Repairmen at Chanute Air Force Base (Dallman, 1977). The 37 CAI lessons in the program, comprising almost 30 hours of instruction and 37 tests, are implemented on the PLATO system along with a routing program that provides individualized instructional management. The 37 lessons are homogeneous in subject matter and tutorial in style for the most part. They are arranged in mastery learning fashion, so that students must achieve the mastery level of the test which was given at the end of each lesson in order to be advanced

Table 1

Summary of Master Validation Exams in the Chanute PLATO AFB CBE Project

Lessons	M[a]	Validation Date	Size of tested out sample	% of Success	% of Failure	Total N	# of Success
103	30	10 June	63	89%	11%	93	83
104a	30	14 April	114	94%	6%	144	134
104b	30	14 April	113	86%	14%	143	124
105	30	14 April	102	88%	12%	132	117
106	30	19 June	33	82%	18%	63	54
201a	30	28 May	99	90%	10%	129	116
201b	30	23 May	109	72%	28%	139	105
202a	30	18 Aug	33	82%	18%	63	54
202b	30	28 May	90	98%	2%	120	115
203a	30	28 May	33	97%	3%	63	59
203b	30	13 June	33	94%	6%	63	58
203c	30	18 Aug	33	91%	9%	63	57
204	30	18 Aug	33	94%	6%	63	58
205a	30	15 Jan	33	79%	21%	63	53
205b	30	15 Jan	33	82%	18%	63	54
206a	30	13 June	90	82%	18%	120	101
206b	30	25 June	65	82%	18%	95	80
206c	30	11 April	118	95%	5%	148	139
207	30	15 Aug	33	91%	9%	63	57
301	30	25 June	109	79%	21%	139	113
304	30	25 June	65	82%	18%	95	80
305	30	18 May	109	96%	4%	139	132
307	30	14 April	130	81%	19%	160	132
308	30	18 May	109	63%	37%	139	96
401	30	17 April	142	83%	17%	172	146
402	30	8 July	65	79%	21%	95	78
403	30	30 June	65	79%	21%	95	78
404	30	2 Sept	33	100%	0%	63	60

[a]M is the sample size used for establishing validation dates.

(Table 1 cont.)

Lessons	M^a	Validation Date	Size of tested out sample	%% of Success	% of Failure	Total N	# of Success
405a	30	26 Aug	33	100%	0%	63	60
405b	30	26 Aug	33	91%	9%	63	57
405c	30	26 Aug	33	94%	6%	63	58
405d	30	2 Sept	33	73%	27%	63	51
406	30	30 June	65	95%	5%	95	89
407	30	22 Sept	33	88%	12%	63	56

to the next lesson. If the mastery level is not achieved, the student must repeat the lesson. The 37 tests consist mostly of matching and multiple-choice items. Mastery levels are aimed at 80% level, but the actually used cutoff are somewhere between 75% and 90% of the items answered correctly. Test lengths vary from 5 to 20 items and the scores on the first try of each item are summed to yield the total score of each test. The tests are called MVE, for Master Validation Exams. For example, the test at the end of lesson 101 is called MVE101. The description of their lessons is given in Appendix2.

A lesson is said to be validated when 90% of the students have achieved the given criterion level of 75% - 90% of the items answered correctly in the first attempt on each master validation exam. The sample consisted of about 30 students from successive clsses. No major modifications of lessons were made until all students in the sample finished the lessons. All lessons were validated according to this criterion between April and September of 1975. The exact validation dates of the lessons are shown in Table 1. In order to validate the validation criterion, the lessons that were said to be validated were left unchanged during the evaluation period and were tested on more students who came in after the validation dates were established.

It is interesting to note that only 15 out of 34*lessons achieved the criterion level of 90% success rate at the end of the evaluation period, although all lessons are labeled "validated." Indeed, this result can be expected and is not very suprising. The next sections will be devoted for explaining the reason.

*The lessons available for the analysis was reduced to 34 from 37.

2.3 Bayesian-Binomial Model

By applying a sample binomial model to the first 30 subjects with whom the validation dates were established, we obtain the result that the probability of failure to meet the validation criterion upon follow-up testing is 36.3 % . Therefore, 12 out of 34 lessons are predicted to be failures. Similarly, the posterior distribution of Bayesian binimial model where beta function was taken as a prior distribution predicts 59.1% failure to meet the validation criterion (this calculation was done by the PLATO version of CADA developed by Mel Novick). In other words, 20 out of 34 lessons are predicted to miss the validation criterion. Table 1 shows that 19 lessons havea failure rate greater than 10%, which is very close to the number (20) predicted by the Bayesian binomial model. This fact indicates that it is necessary to introduce a more accurate validation criterion for lessons. The reader might wander how the prior distribution was chosen here. It was based on the belief of the people who participated the PLATO AFB CBE project.

Producing a lesson to be used on the PLATO system is not a simple task. Many steps are involved in the completion of a lesson, including tryout with students and gathering empirical evidence which might indicate further revision or modification of the lessons. No unique method for lesson-revision operation, based on the theories of educational psychology and educational measurement, has been developed for use on the PLATO system. As signals pointing to the need for revision, some authors choose to look at "Area Data," which is collected by the computer, and consists of elapsed time in the area (a segment of instruction), number of questions answered correctly on the first try (Okf's), number of incorrect responses to questions (no's), number of correct responses to questions (Ok's), and number of helps requested. Others design and implement their own data collection routines. These data usually give lesson authors a very rough idea of the how well their lessons work with students and indicate the areas where the majority of students had trouble going through.

Thus, it is possible for a PLATO lesson author to have some degree of confidence in the quality of his lessons by the time the lesson becomes a nearly finished product. The degree of his confidence might depend on his knowledge of teaching strategies or his past experience. If he uses teaching strategies such as mastery learning, which has been examined by many researchers and is known to be highly effective, then it is natural to assume that he would be highly confident of the quality of his lesson. If an author has substantial experience producing lessons on the PLATO system and has used them successfully in his class, then his experience will assure him of the success of his new lesson.

It must be true that lessons in which the author has high confidence are more likely to produce a higher success rate in a future use of his lessons. Suppose p is the true probability of success associated with a given lesson; in other words, p % of students achieve the mastery level in a population. In general, a Bayesian density indicates a state of belief about a parameter, such as p here, intermediate between the estimate "I know nothing about p" and "I know the exact value of p."

Two types of densities are used, one being the prior density, representing beliefs about the parameter before observations are obtained, and the other being the posterior density, reprensenting beliefs after seeing the data. In our situation, the task is to infer the value of p from an observation x. It is clear that p obtained in this way cannot be exact: that 20 students passed the test out of 25 students is quite a probable number for lessons with the value of p anywhere between .65 and .90. But the observation that 80% of students achieved the mastery level makes p around .8 more likely for the lesson than p around .3, so we should estimate p as .8 if nothing else is known about the quality of the lessons. If the author has some information about the lesson, such as that since the lesson is dealing with a simple introductory task, the value of .8 is somewhat lower than it should be, then we would be more inclined to think that the true probability of success associated with the lesson is higher than .8. If the author has substantial experience in producing high quality lessons in past years, then his new lesson would be more likely to be considered to have a higher true probability of success than .8, even though the observed success rate is .8 in the sample. Therefore, our estimate of the true probability p depends not only on the observed value x, but also on what we know about p before observing x.

The previous knowledge can be expressed by a prior density function $f(p)$ (or, also called a prior probability density function). The product of $f(p)$ and the likelihood function $f(x|p)$ (i.e., the conditional probability of x on given p) gives a quantity proportional to the posterior density function $f(p|x)$:

$$f(p|x) = f(x|p)f(p)$$

where $f(x|p)$ is called the model density function instead of likelihood, as in Bayesian statistics.

The model density is used for inference in traditional statistics, or sampling theory. It is clear that Bayesian statistics uses more information than traditional statistics does, i.e.,the prior density function. Consequently, Bayesian statistics will provide us with more accurate information, at least mathematically, than traditional statistics will if a choice of our prior density is the right one. Indeed, it is possible to demonstrate such an example, especially if the number of observations is fairly small. But it is true that the model density, conditional probability of x given p, will have most influence on the posterior density when the number of observations is large.

A detailed discussion of Bayesian binomial model can be found elsewhere (Novick and Jackson,1975; Ferguson, T., 1971). We will show only the Bayesian densities in this paper. If we assume the prior belif of p follows a beta distribution, then the prior density f(p) is given by a beta function:

$$f(p) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)} \quad , \quad 0 \le p \le 1, \ a>0, \ b>0$$

the model density $f(x|p)$ is

$$f(x|p) = \frac{p^{x-1}(1-p)^{N-x}}{B(x,N-x-1)}$$

the posterior density $f(p|x)$ is given by

$$f(p|x) = \frac{p^{a+x-1}(1-p)^{b+N-x}}{B(a+x, b+N-x)}$$

where $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$, N is the number of subjects.

Application of the Bayesian binomial model to 34 Chanute lessons will be demonstrated in the next section.

2.4 Appropriateness of the Percentage of Success Rate

The rule for establishing validation of a lesson was that 27 of 30 students entering the lesson successively must pass the mastery test given at the end of the lesson; if this criterion was not met, some revision of the lesson was carried out. If we consider the 34 lessons are homogeneous, as Dallman (1977) stated in his paper, the model density function derived from a sample of size 30 with 27 successful attempts predicts a 63.7% chance of success for each lesson in future at the time when the validation date was established.

The corresponding prior density in our situation is obtained from the validation criterion (which has been used in CBE programs in the Army (Branson et al. 1975): 27 of 30 achieving criterion level. It was believed that this rule was adequate to determine the cutoff point for terminating the process of lesson modification and beginning to gather data for evaluating the PLATO AFB CBE project at Chanute. The belief that a 90% rate of success in thirty successive subjects is an adequate criterion for validating lessons, can be thought of as the prior condition. Therefore, the same beta-binomial distribution function as the model density function is taken as a prior density distribution in this case.

Applying Bayes' theorem to prior and model densities, the posterior density function is given by beta-binomial function B(53.2, 6.8) with a mode of .87 and standard deviation of .04. The 50% credibility interval is given by [.8714, .9244], in which mode .9 and mean .87 are included.

In Bayesian statistics, the interval [.8714,.9244] is called a 50% credibility interval for the ability (or success rate) because the 50% is the measure of the strength of our belief, taking into account our prior knowledge and our observation that the student's (or lesson's) ability lies in that interval. In particular [.87, .92] is a 50% interval between the 25th and 75th percentiles and is called the highest-density region in the belief, a 50% HDR. The length of the interval .92 - .87 is called an interquartile range and is used as a measure of variability of distribution.

As seen in Table 1, we have further observations made after the validation dates were established. Let us extend our discussion further.

Table 2 summarizes the results of the Bayesian beta-binomial analysis for each lesson based on the expanded sample and newly observed success rate. The model density functions of the lessons given in Table 2 were derived from the new sample of size given in column 8 and number of successes in column 9 of Table 2. The parameters of prior density, 50% HDR and probabilities of π larger than or equal to .9 (Prob($\pi \geq .9$), are given in Table 2. From the last column of Table 2 we may select the lessons whose probabilities of being validated lessons are greater than .50. Since all standard deviations and interquatile ranges are small, i.e., mostly less than .05, the probability that π is greater than or equal to .85 will be drastically greater.

For example, lesson 105 has Prob($\pi \geq .85$)=.86 while Prob($\pi \geq .9$) = .25. Therefore, it is recommended that the validation criterion of 90% be replaced by a slightly higher value 92% or so. If we defined the validaton criterion by a slightly higher success rate, say, 28 out of 30 students

Table 2

Credibility Intervals of Master Validation Exams

by Baysian Binomial Model

Lessons	Observed Score	Mean	Mode	S.D.	a	b	50% CI	$P(\pi \geq .90)$
103	$\frac{83}{93} = .892$	.89	.89	.03	109.2	13.8	.8744, .9120	.36
104a	$\frac{134}{144} = .931$	.93	.93	.02	133.2	10.8	.9157, .9444	.87
104b	$\frac{124}{143} = .867$	.87	.86	.03	123.2	19.8	.8467, .8851	.08
105	$\frac{117}{132} = .886$	.89	.88	.03	116.2	15.2	.8665, .9040	.25
106	$\frac{54}{63} = .857$	.86	.84	.05	53.2	9.8	.8238, .8842	.10
201a	$\frac{116}{129} = .899$	.90	.89	.03	115.2	13.8	.8800, .9160	.43
201b	$\frac{105}{139} = .755$	.75	.75	.04	104.2	34.8	.7280, .7774	.00
202a	$\frac{54}{63} = .857$	.86	.84	.05	53.2	9.8	.8238, .8842	.10
202b	$\frac{115}{120} = .958$	.95	.94	.02	141.2	8.8	.9340, .9588	.97
203a	$\frac{59}{63} = .937$	.93	.92	.03	85.2	7.8	.9052, .9425	.74
203b	$\frac{58}{63} = .921$	.92	.91	.04	57.2	5.8	.8959, .9425	.63
203c	$\frac{57}{63} = .905$	.90	.89	.03	83.2	9.8	.8811, .9228	.47
204	$\frac{58}{63} = .921$	.92	.91	.04	57.2	5.8	.8959, .9425	.63
205a	$\frac{53}{63} = .841$	.86	.85	.04	79.2	13.8	.8337, .8826	.08
205b	$\frac{54}{63} = .857$	.86	.84	.05	53.2	9.8	.8238, .8842	.10
206a	$\frac{101}{120} = .842$	.85	.85	.03	127.2	22.8	.8324, .8716	.97
206b	$\frac{80}{95} = .842$	.86	.85	.03	106.2	18.0	.8331, .8758	.05

(Table 2 cont'd)

Lessons	Observed Score	Mean	Mode	S.D.	a	b	50% CI	$P(\pi>.90)$
206c	$\frac{139}{148} = .939$	.94	.93	.02	138.2	9.8	.9255, .9521	.94
207	$\frac{57}{63} = .905$	.90	.89	.03	83.2	9.8	.8811, .9228	.47
301	$\frac{113}{139} = .813$	.83	.82	.03	139.2	29.8	.8073, .8466	.00
304	$\frac{80}{95} = .842$	.86	.85	.03	106.2	18.8	.8331, .8758	.04
305	$\frac{132}{139} = .950$	.94	.94	.02	158.2	10.8	.9282, .9528	.96
307	$\frac{132}{160} = .826$	.84	.83	.03	158.2	31.8	.8175, .8538	.00
308	$\frac{96}{139} = .691$	.73	.72	.03	1222.0	46.8	.7020, .7485	.00
401	$\frac{146}{172} = .849$	.86	.85	.03	172.2	29.8	.8380, .872	.00
402	$\frac{78}{95} = .821$	.84	.83	.03	104.2	20.8	.8160, .8604	.013
403	$\frac{78}{95} = .821$	.84	.83	.03	104.2	20.8	.8160, .8604	.013
404	$\frac{60}{63} = .952$	.94	.93	.03	86.2	6.8	.9174, .9522	.84
405a	$\frac{60}{63} = .952$	.94	.93	.03	86.2	6.8	.9174, .9522	.84
405b	$\frac{57}{63} = .905$	.90	.89	.03	83.2	9.8	.8811, .9228	.47
405c	$\frac{58}{63} = .921$	.92	.91	.04	57.2	15.8	.8959, .9425	.63
405d	$\frac{51}{63} = .810$	.84	.83	.04	77.2	15.8	.8103, .8622	.02
406	$\frac{89}{95} = .937$	.93	.92	.02	115.2	9.8	.9117, .9431	.82
407	$\frac{56}{63} = .889$	.89	.88	.04	55.2	7.8	.8595, .9137	.31

achieving the mastery level in a successive sample, then the validation dates given in column 4 of Table 2 $p(\pi > .9)$ would be later dates but the estimation of true probability of success would be much improved. Lesson 201a has a 90% success rate in an observation of 99 students who entered the lesson after the validation date, May 28th. This observed success rate is the same as the validation criterion. It is interesting to note that the 50% HDR [.88,.916] of the new prior density based on the sample size of 129 is slightly narrower than that of size 30 [.8714, .9244]. In general, when the number of students increases , the 50% HDR gets narrower. Also you will notice that the value in the last column of Table 2 for lesson 201a is .43, which is larger than $Prob(\pi \geq .9)$ = .409 when the sample size is 30. Therefore, our crediblity of saying that lesson 201a will have a success rate of 90% in the population from which this sample was drawn will increase if the sample size on which the model density was based increases.

Hence, setting the most appropriate validation criterion for a lesson depends on two factors: success rate and sample size. The discussion of these two factors will be carried mathematically parallel, ,in other words mathematically dual; taking the sample size as the number of items or the test length, the success rate as the proportion of getting a correct answer for an item. In the next chapter, we will switch the focus from the former that is oriented toward the success rate of a lesson,to the latter that is for the success rate of an individual in a test.

CRITERION REFERENCED TEST AS ASSESSMENT OF STUDENTS PERFORMANCE

3.1 Problems in Criterion-Referenced Tests

Criterion-referenced testing has gained much attention from educational measurement and testing specialists in recent years. The object of criterion-referenced testing is not to distinguish finely among subjects, but to classify subjects into mastery and non-mastery groups. Robert Gleser (1963) stated that the measures of CRTs depend on an absolute standard of quality while those of NRTs depend on a relative standard. CRTs are often used in conjunction with instructional programs that maximize the number of students attaining a given mastery level and minimize the variability of test scores while norm-referenced tests (NRTs) are used in selection or screening a subgroup of examinees, predicting students' future performances, and evaluation of instructional programs.

The concepts of criterion-referenced testing are quite different from those of norm-referenced testing. Strictly speaking, the test scores of NRT are assumed to be distributed normally while those of a CRT are highly skewed. The variability in scores of a NRT is large while that of a CRT is small. Although, these differences are generally expected but need not be observed in practice. Statistical measures in the classical test theory model, such as reliability and validity, are defined on the basis of assuming that the standard deviation of any NRT is always positive and adequately large. Therefore, the definition of reliability as the ratio of true score variance to observed score variance can be a meaningful index there. The reliability tends to increase as the test length (number of items) increases and hence the variablity of test scores increases. The test length of a CRT is usually short, say 10 or 15 items, and often most items of a test are answered correctly by all students who take the test. Therefore the reliability of a CRT can't be satisfactorily large. As far as this author knows, many tests have a α_{21} reliability of only about 0.5 or less.

Since it is a common use of criterion-referenced testing that all students are expected to achieve the level of mastery, say 90% correct, the observed scores become a bounded variable. If there are subjects with true scores near the "ceiling" or the "floor", it becomes implausible to assume that the errors of measurement are distributed independently of true scores for those near the boundary. NRTs don't usually have ceiling or floor effects. Their scores are distributed around the mean score and are seldom near either extreme. In such a test, it is reasonable to assume that error scores are due to something independent of the subject's true abilities, such as fatigue, anxiety, etc.

Lord and Novick (1968) argue about the plausible distributional forms of observed CRT scores and true scores in Chapter 23 of their book, "Statistical Theories of Mental Test Scores." We will follow their steps and adopt the binomial error model for CRT scores. The binomial error model assumes that if each MVE test is aimed at measuring the learning level of a topic taught in the Vehicle Training Course, then all items in the test must measure the same task. In other words all items in a test have one and only one common factor with 0-1 scoring. Suppose there is a pool of items measuring the same task, and taking an item out of the pool is an independent event, that is, answering the earlier items on the test does not affect the ability of a student to answer later items correctly, then we can formulate the distribution of raw scores x by a binomial distribution with parameter θ in which θ is the proportion of items that a student would answer correctly over the entire pool of items. If T is a fixed true score and e is an error of measurement, then the raw score x can be expressed by the sum of the two, $x = T + e$, and θ is given by

$$\theta = T/n$$

where n is the number of items in the test. Let $h(x|\theta)$ be the binomial distribution of x at any given true ability level θ, then the conditional distribution $h(x|\theta)$ can be given by

$$h(x|\theta)=\binom{n}{x}\theta^{x}(1-\theta)^{n-x} \qquad x = 0,1,\ldots,n.$$

where n is the number of items in the test.

It is interesting to note that this model does not pay attention to item differnces. The traditional measurement indices such as item difficulty or items discriminating index are not the major concern in the binomial error model. Rather, finding out how accurately a test can estimate an examinee's pass or fail status with respect to a given mastery is a main concern of the model.

Keats and Lord (1962) investigated the relationship between the distibution of test scores, observed and true scores. The test scores could be adequately represented by the hyper geometric distibution h(x) with a negative parameter and the true scores distribution could be represented by the two parameter beta distribution $g(\theta)$.

$$g(\theta) = \theta^{a-1}(1-\theta)^{b-n} / B(a,b-n+1)$$

where $a>0$ and $b>n-1$. And also

$$h(x) = \int_0^1 \frac{\theta^{a-1}(1-\theta)^{b-n}}{B(a,b-n+1)} \binom{n}{x} \theta^x (1-\theta)^{n-x} d\theta \ , \quad x=0,1,\ldots,n.$$

In classical test theory, the estimation of a true score is given by regressing the true score T on the observed score x, and the equation is given by

$$E(T|x) = \rho x + (1-\rho)u_x$$

where ρ is the reliability of the test and u_x is the mean of test scores.

In binomial error model, the estimation of a true score is given by similar equation,

$$E(T|x) = \alpha_{21}x + (1-\alpha_{21})u_x \ , \quad x=0,1,\ldots,n$$

where α_{21} is the ratio of number-correct true-score variance to observed-score variance and is given by

$$\frac{\sigma^2_T}{\sigma^2_x} = \frac{n}{n-1}\left\{1- \frac{u_x(n-u_x)}{n\sigma^2_x}\right\} = \alpha_{21}$$

Table 3 is the summary of information from the Mastery Validation Exams at Chanute.

Table 3

The Summary of Simple Statistics of Mastery Validation Exams

test	mean	SD	items	α21	N
mve103	7.388	1.124	8	0.6321	85
mve104a	11.892	0.442	12	0.4910	83
mve104b	10.120	1.728	11	0.8018	83
mve105	7.706	0.737	8	0.5470	85
mve201a	9.474	0.973	10	0.5254	76
mve201b	8.907	1.325	10	0.4951	86
mve202a	16.186	2.934	20	0.6753	97
mve202b	9.720	0.634	10	0.3573	82
mve204	8.557	1.681	10	0.6253	88
mve205a	6.767	1.558	9	0.3470	90
mve205b	8.110	1.736	10	0.5457	82
mve206a	12.038	1.574	13	0.6942	78
mve206b	15.250	1.619	17	0.4259	80
mve206c	19.257	1.151	20	0.4841	70
mve207	3.761	1.124	5	0.3287	88
mve301	8.727	1.501	10	0.5635	77
mve303	17.380	2.257	20	0.5824	71
mve304	9.209	1.366	10	0.6771	67
mve305	7.458	0.934	8	0.4806	72
mve307	14.683	1.522	16	0.5101	63
mve308	9.037	1.170	10	0.4045	82
mve401	9.254	1.015	10	0.3673	63
mve402	14.138	2.335	17	0.5988	94
mve403	8.095	2.487	10	0.8340	84
mve404	4.254	0.876	5	0.2166	67
mve405a	9.169	1.069	10	0.3701	71
mve405b	8.329	1.991	10	0.7208	70
mve405c	9.087	1.222	10	0.4934	69

In classical test theory, α_{21} (Kuder-Richardson)) is always smaller than or equal to the other reliabilty approximations ,such as α_{20} and Cronbach's coefficient α. Both α_{20} and α_{21} become equal only when all items are of equal difficulty (or have equal mean if the scores are dichotomous, and note that α_{20} would be used in place of α_{21} with a compound binomial model). Coefficient α becomes equal to α_{20} if all items in a test are parallel, that is, all items have the same mean values and variances in classical test theory. As we previously noted in this chapter, the binomial error model assumes a single common factor and is not concerned with differenciating among item characteristics. The model does not require any information about the item charactersitics in a test, such as difficulty and discriminating index, but it does require knowledge of the number of items on a test. It is interesting to note that the mathematically derived ratio of the true and observed score variances in the model becomes equal to the reliability of the test where all items are of equal difficulty and variance. Therefore the definition of reliability in classical test theory loses an interesting feature in terms of a traditional sense because in the binomial error model, the value of the reliability index is reduced to that of the lowest approximation to the ratio of the true and observed score variances in classical test theory. Since α_{21} is a special case of reliability approximations when item differences are ignored, it is exactly what we can expect out of the binomial model.

The conceptualization of reliability is no longer important in the model. Instead, the accuracy of judging non-mastery and mastery status of examinees becomes a main concern. Millman states this purpose of CRT clearly in his paper (1975), and discusses how many items must be administered from a given item-pool so that the test items in the domain answered correctly can give an accurate estimation of an examinee's true ability θ.

Setting of Mastery Levels

The mastery level of Master Validation Exams (MVE) of the 37 lessons in the Chanute PLATO AFB CBE Program was set at a level of 80%, although it is impossible to prove that 80% is the most appropiate level for their program. Block (1972) showed in his experimental study that attainment of a 95% mastery level maximized student learning of cognitive tasks in his matrix algebra course, while an 85% level maximized learning as characterized by affective criteria.

Since Chanute's 37 lessons are designed to be "homogeneous" with respect to content and teaching style, all lessons are written under the same principle with the same tutorial logic, although the subject matter in each lesson is different. Therefore Chanute's lessons are not linearly related and the content difficulty of the lessons is not hierarchically ordered as it would be in teaching mathematics, arithmetic, or foreign languages. If the lessons are linearly related, setting a mastery level for the earlier instructional units should by higher than those of the later instructional units. If the goal of the second unit is the attainment an 85% mastery level, then the mastery level of the first unit might be 90%, or some other level higher than 85%. Since there is no analytical technique to provide the optimal level of mastery learning, definite statments about the determination of ideal mastery levels cannot be made at this time. Linn (1978) provides an excellent discussion of the topic of "setting standards".

Cutoffs

Mastery levels are usually set by instructors or the author of a lesson, but the decision of mastery and non-mastery is based on examinees' observed test scores. The score that is used as to decide mastery and non-mastery is called the "cutoff." Mastery and non-mastery status ought to be defined on the basis of true ability θ, not observed test scores x that are subject to measurement errors. If true ability were known, there would be no incorrect classifications. Unfortunately, true scores are impossible to obtain in practice, so we have to find a way to minimize misclassification.

There are four kinds of classifications: 1. an examinee's true ability θ and observed score x are both higher than a given mastery level θ_o and cutoff score c, that is A = { $x \geq c$ and $\theta \geq \theta_o$ }; 2. θ is lower than θ_o and x is also lower than c, that is B = { $x < c$ and $\theta < \theta_o$ };3. θ is lower than θ_o, but x is larger than c, F_+ = { $x \geq c$ and $\theta < \theta_o$ }; 4. θ is higher than θ_o, but x is lower than c, F_- = { $x < c$ and $\theta \geq \theta_o$ }. The following figure shows these four conditions.

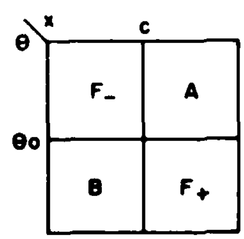


Figure 1

θ = true ability, x = observed score

θ_o= true mastery level

c = observed cutoff

Probability of these events will be denoted by $P(A)$,$P(B)$,$P(F_+)$ and $P(F_-)$ respectively

Millman (1975), and then Novick & Lewis (1975) reported percent of students expected to be misclassified for a given cutoff with various numbers of test items. Millman used the binomial error model, but Novick and Lewis used the Bayesian beta binomial error model.

According to Millman's calculations, the percent of students expected to be misclassified at 80% mastery level using a 10 item test could be as high as 53%.

Emerick (1972) and Huynh (1976) considered the loss ratio Q of F- to F+ as a means of controlling misclassification, especially false advancement. If later instructional units require the knowledge and skill acquired in earlier units, false advancement will be a problem. Since F- stands for the event in which a student has really mastered the given instructional unit but his/her observed score happens to be lower than the cutoff, retaining such a student in the same unit is not efficient. If the instructional units are fairly independent from one to another, as are lessons in the Vehicle Training Program at Chanute Air Force Base, then an appropriate loss ratio would be 1, or at least it is not necessary to set it as high as 10.

Huynh (1976) proposed an evaluation of the cutoff score that minimizes the occurence of misclassifications for a given loss ratio. With his cutoff score, the loss ratio associated with the probability of having the false positive to that of false negative stays the same, say 10, while the linear combination of the probabilities of the both events and the loss ratio (the average loss) is minimized. We will discuss in more detail Huynh's method in conjunction with 34 Chanute lessons and their MVE test scores.

3.2 Evaluation of the optimal cutoff scores

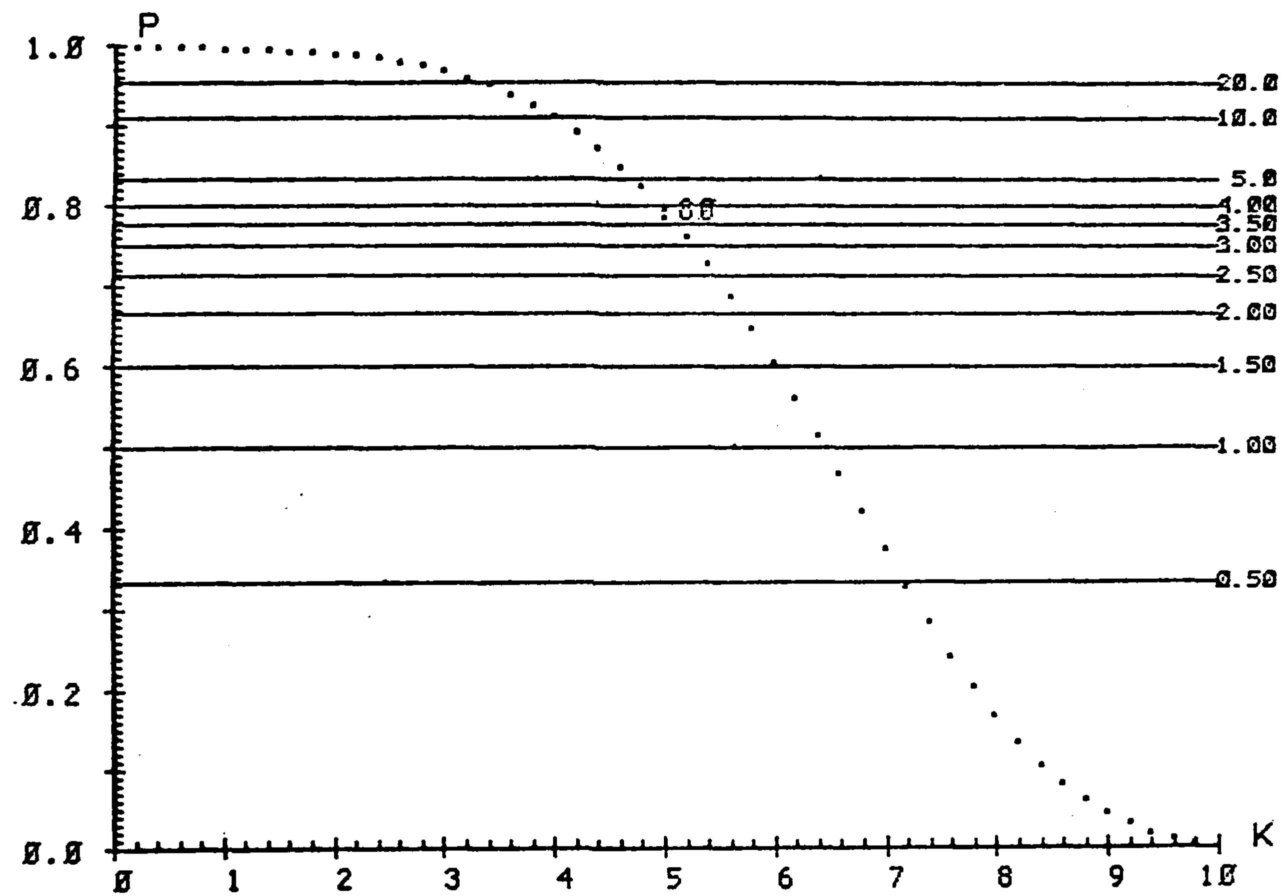
Huynh derived the optimal cutoff c_o of a test for a given mastery level θ_o and loss ratio Q so as to minimize the average loss function R(c) by differentiating it, where R(c) is the linear combination of the probability of false positive and false negative and is given by

$$R(c) = P(F+) + Q\ P(F-).$$

c_o is the smallest integer such that the incomplete beta function of $I_{\theta_o}(a+c_o, n+b-c_o)$ is smaller than or equal to $Q/(1+Q)$; where

$$p(c_o) = I_{\theta}(a+c_o,\ n+b-c_o) = \int_0^{\theta_o} \frac{\theta^{a+c_o-1}(1-\theta)^{n+b-c_o-1}}{B(a+c_o,\ n+b-c_o)}\, d\theta$$

In order to apply Huynh's result to evaluate c_o, we need the help of a computer to calculate the values of the incomplete beta function for c=0,1,2,...n and plot them on paper. The PLATO system eases these steps and we can obtain the answer through the program "cutoff" written by the present author and T. Weaver. Figure 2 illustrates the procedure to determine the optimal cutoff c_o. The parameters a and b are obtained from the mean, standard deviation of the test and the number of items in the test (denoted by n). Table 4 shows the values of incomplete beta function $I_{\theta_o}(i)$ at each point i=1,2,... n, where a,b are calculated from test scores of MVE201a by the formula,



NEXT for ts=Ø.85 OR type ts ≫
HELP to calculate p by k and ts

Figure 2

Determining the optimal cutoff C_0 as to minimize misclassification

lesson = MVE201a	subjects = 76	n = 10
mean = 9.4737	SD = 0.9726	x_{21} = 0.53
a = 8.5560	b = 0.4753	

$a = (-1+1/21)\ u_x$

$b = -a-n+n/21.$

Table 4

Ten points in Figure 2

$\theta_o = .80$, Test=mve201a, a=8.5560 b=0.4753

item	a+i	n+b-i	$I_{\theta_o}(a+i,n+b-i)$
1	9.556	9.475	0.998
2	10.556	8.475	0.991
3	11.556	7.475	0.969
4	12.556	6.475	0.913
5	13.556	5.475	0.796
6	14.556	4.475	0.608
7	15.556	3.475	0.376
8	16.556	2.475	0.169
9	17.556	1.475	0.045
10	18.556	0.475	0.004

The curve in Figure 2 is obtained by plotting the points in Table 4. The horizontal lines which are marked by losses 0.5, 1, 1.5, 2,...,20 in Figure 2 help to evaluate the optimal cutoff which minimizes the average loss R(c) at c_o for the partially known loss ratio Q and a given mastery true level θ_o. Since the contents of all lessons discussed in the Chanute PLATO AFB CBE Program deal with independent topics across the lessons and the lessons are not linearly or hierarchically related, a loss ratio of 1 will be reasonable. Thus, in Figure 2 the smallest integer value of i for which the curve P(i) goes under the line of loss ratio 1 is 7. Therefore c_o=7 is the ideal cutoff score of the test, MVE201a.

It is interesting to note that the cutoff score, c=8, actually used for MVE201a in the Chanute training program gives a slightly larger value of the probability of misclassification of $(R(c)=P(F+)+P(F-))$ than the theoretically derived c_o does, but not for P(F+), probability of false positive, or P(F-), probability of false negative separately.

$$P(F+) = I_{\theta_o}(a,b)-(1/B(a,b))\sum_{i=0}^{c-1}\binom{n}{i}B(a+i,b+n-i)I_{\theta_o}(a+i,b+n-i)$$

$$P(F_-) = (1/B(a,b))\sum_{i=0}^{c-1}\binom{n}{i}B(a+i,n+b-i)(1-I_{\theta_o}(a+i,b+n-i))$$

The probabilities of $P(A)=\text{Prob}\{\theta\geq\theta_o,x\geq c\}$ and $P(B)=\text{Prob}\{\theta<\theta_o,x<c\}$ are given respectively by the following formulas:

$$P(A) = 1-I_{\theta_o}(a,b)+(1/B(a,b))\sum_{i=0}^{c-1}\binom{n}{i}B(a+i,n+b-i)(I_{\theta_o}(a+i,b+n-i)-1)$$

$$P(B) = (1/B(a,b))\sum_{i=0}^{c-1}\binom{n}{i}B(a+i,b+n-i)I_{\theta_o}(a+i,b+n-i)$$

The probability of each misclassification for all available MVEs were calculated and summerized in Table 5.

Since the sum of the probabilities A, B, F+, and F- is 1, the sum of the probabilities of A and B must have a maximum value at c_o where P(F+)+P(F-) reaches the minimum as shown in Figure 3.

In Figure 3, the curve of P(F+)+ P(F-)(the lower curve drawn * is) decreases slowly until it reaches the bottom at c_o, then increases as the number of items increases while the curve of P(A)+ P(B)(the upper curve drawn with + is) reaches the maximum point at c_o.

Table 5

Estimated Probability of Misclassifications

Test	Cutoff[a]		$P(F_+)$	$P(F_-)$	$P(F_+$ or $F_-)$	$P(A$ or $F_+)$	Success rate
mve103	c_O	6	0.0621	0.0162	0.0783	0.9247	.89
	c	7	0.0314	0.0639	0.0953	0.8462	
mve104a	c_O	7	0.0026	0.0001	0.0026	0.9997	.94
	c	10	0.0011	0.0057	0.0068	0.9927	
mve104b	c_O	9	0.0348	0.0259	0.0606	0.8705	.86
	c	9	0.0348	0.0259	0.0606	0.8705	
mve105	c_O	6	0.0235	0.0094	0.0329	0.9739	.88
	c	7	0.0123	0.0399	0.0522	0.9323	
mve201a	c_O	7	0.0357	0.0064	0.0421	0.9788	.90
	c	8	0.0238	0.0262	0.0499	0.9472	
mve201b	c_O	7	0.1078	0.0146	0.1223	0.9375	.72
	c	8	0.0710	0.0556	0.1266	0.8598	
mve202a	c_O	16	0.1163	0.0624	0.1788	0.6495	.82
	c	16	0.1163	0.0624	0.1788	0.6495	
mve202b	c_O	5	0.0055	0.0001	0.0056	0.9998	.98
	c	8	0.0031	0.0122	0.0153	0.9853	
mve204	c_O	8	0.0996	0.0503	0.1499	0.7803	.94
	c	8	0.0996	0.0503	0.1499	0.7803	
mve205a	c_O	8	0.1428	0.1341	0.2769	0.3612	.79
	c	8	0.1428	0.1341	0.2769	0.3612	
mve205b	c_O	8	0.1507	0.0634	0.2141	0.6913	.82
	c	8	0.1507	0.0634	0.2141	0.6913	
mve206a	c_O	10	0.0478	0.0184	0.0662	0.9207	.82
	c	11	0.0266	0.0535	0.0801	0.8644	
mve206b	c_O	12	0.0606	0.0113	0.0719	0.9708	.82
	c	14	0.0305	0.0911	0.1216	0.8608	
mve206c	c_O	13	0.0057	0.0003	0.0061	0.9991	.95
	c	16	0.0030	0.0116	0.0146	0.9852	
mve207	c_O	5	0.0965	0.1957	0.2922	0.3070	.91
	c	4	0.2878	0.0547	0.3425	0.6393	

Table 5 (cont.)

	Cutoff[a]		$P(F_+)$	$P(F_-)$	$P(F_+ \text{ or } F_-)$	$P(A \text{ or } F_+)$	Success rate
mve301	c_o	8	0.0894	0.0540	0.1434	0.8184	.79
	c	8	0.0894	0.0540	0.1434	0.8184	
mve303	c_o	15	0.1070	0.0266	0.1336	0.8867	.90
	c	16	0.0730	0.0653	0.1383	0.8140	
mve304	c_o	8	0.0471	0.0292	0.0763	0.8922	.82
	c	8	0.0471	0.0292	0.0763	0.8922	
mve305	c_o	5	0.0632	0.0036	0.0668	0.9827	.96
	c	7	0.0247	0.0787	0.1034	0.8691	
mve307	c_o	11	0.0526	0.0056	0.0582	0.9797	.81
	c	12	0.0413	0.0187	0.0600	0.9553	
mve308	c_o	7	0.0732	0.0147	0.0880	0.9601	.63
	c	8	0.0498	0.0578	0.1076	0.8936	
mve401	c_o	7	0.0364	0.0109	0.0473	0.9872	.83
	c	8	0.0252	0.0451	0.0704	0.9328	
mve402	c_o	13	0.1494	0.0395	0.1890	0.7809	.79
	c	14	0.0910	0.0961	0.1871	0.6660	
mve403	c_o	8	0.0771	0.0294	0.1065	0.7048	.79
	c	8	0.0771	0.0294	0.1065	0.7048	
mve404	c_o	3	0.2100	0.0130	0.2230	0.9564	1.00
	c	4	0.1455	0.0840	0.2296	0.8208	
mve405a	c_o	6	0.0560	0.0025	0.0585	0.9919	1.00
	c	8	0.0326	0.0513	0.0839	0.9196	
mve405b	c_o	8	0.0987	0.0419	0.1405	0.7344	.91
	c	8	0.0987	0.0419	0.1405	0.7344	
mve405c	c_o	7	0.0794	0.0123	0.0917	0.9543	.94
	c	8	0.0527	0.0478	0.1005	0.8921	

[a] c_o is the theoretically derived cutoff to minimize $P(F_+) + P(F_-)$. c is the cutoff actually used in the PLATO Service Program at Chanute.

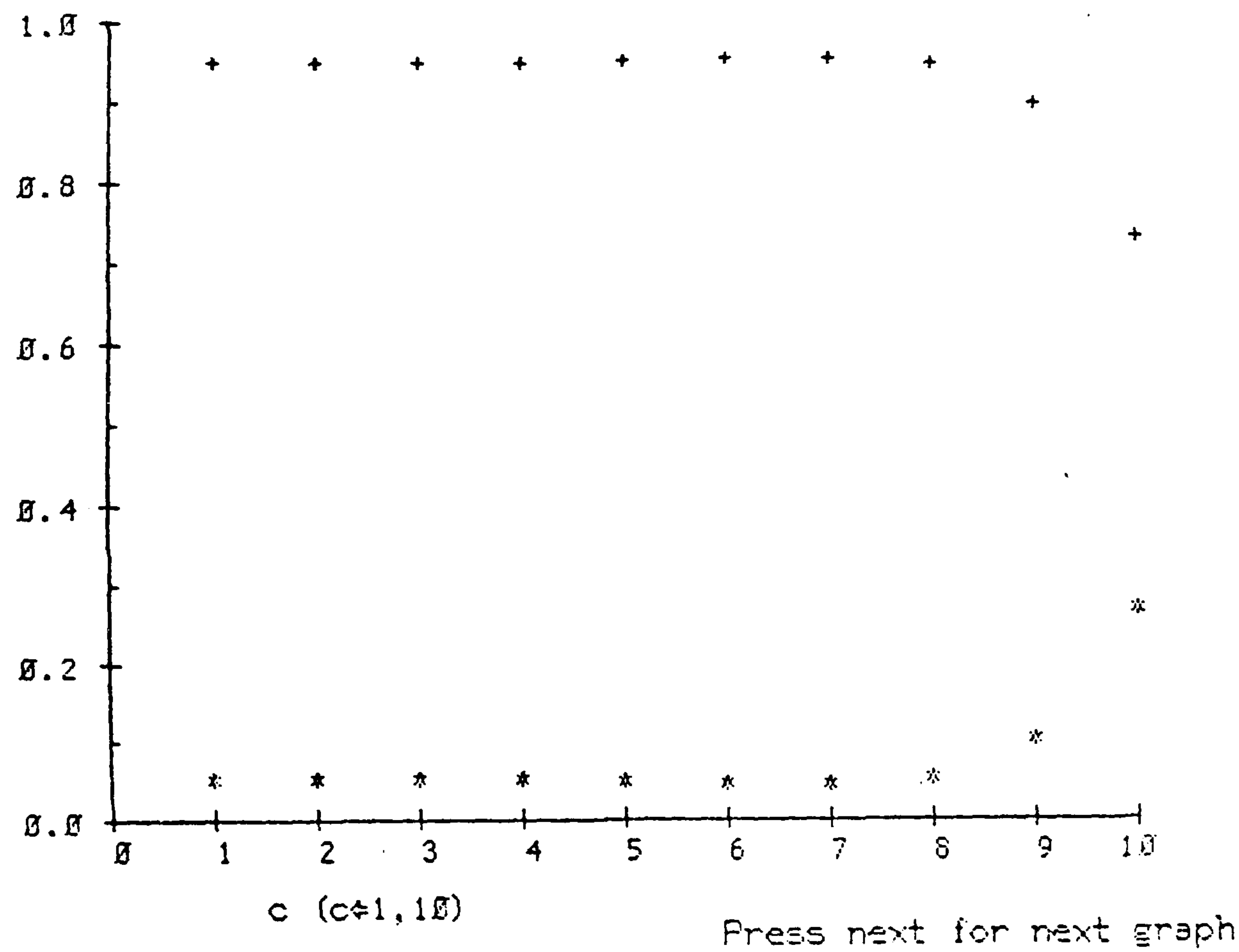


Figure 3

Graph of $P(F_+) + P(F_-)$ over cutoff scores

lesson = MVE201a $n = 10$ $C_0 = 7$ $\theta_0 = .80$

Table 5 indicates that the actually used cutoff scores c produce higher probabilities of P(F+ or F-) than the theoretically determined cutoff c_os except in a few cases. Since the theoretical cutoffs are determined so as to minimize the average loss R(c), in our case the sum of probabilities of false negative F- and false postive F+, all values in column 5 of Table 5, P(F+)+ P(F-) have smaller values for c_o than for c. The sum of the probability of A and F+ is the expected success rate , so this sum matches the observed success rate given in the last column fairly well.

The probability of each misclassification for all available MVEs were calculated and summerized in Table 5.

Since the sum of the probabilities A, B, F+, and F- is 1, the sum of the probabilities of A and B must have a maximum value at c_o where P(F+)+P(F-) reaches the minimum as shown in Figure 3.

In Figure 3, the curve of P(F+)+ P(F-)(the lower curve drawn * is) decreases slowly until it reaches the bottom at c_o, then increases as the number of items increases while the curve of P(A)+ P(B)(the upper curve drawn with + is) reaches the maximum point at c_o.
If c_o were used as cutoffs for MVE test scores, only 12 lessons would not have a probability of observed success less than .90, which was used as the lesson validation criterion in the PLATO AFB CBE program, while 18 lessons have values in P(A)+P(F+)(i.e. $p(x \geq c)$) when c's are used.

Since the probability of false negative, P(F-) stands for the case that an examinee really mastered the goal of instructional unit but his/her observed score happened to be lower than the used cutoff c, he/she does not have to repeat the instruction. If efficiency of training in terms of shortening the training time is the main concern, then P(F-) should not be so large. For example, MVE207 has P(F-)= .1957 which means 88x0.1957=17, out of a total of 88 students repeated the same instruction unnecessarily. Of course this is an extreme case and most p values are less than 10%, which means that five to eight students repeated the same lesson mistakenly. Table 6 shows the number of students misclassified in Master Validation Exams. Since the observed cutoff c for all MVEs but MVE 207 are larger than or equal to the optimum cutoff c_o, the number of misclassified students of the type F+ becomes larger for using c_o than c, and errors in the type F- turn to be smaller for c_o. But the total misclassifications are minimized by using c_o. It is a problem of the tradeoff how the cutoff be selected. Since the loss ratio of 1 was selected in our study, we conclude that most cutoffs of Master Validation Exams used at Chanute were not the best choice. By adopting the theoretically derived cutoff c_o's the probability of misclassifications could have been minimized.

The probabilities of success rate by observation, prob($x \geq c$), or prob(A or F_+), suggest that the validation criterion of lessons in the Chanute program must be changed. Twelve out of 27 lessons have a passing probability of less than .90, even if the theoretical cutoff c_o had been adopted instead of the actually used cutoff score c. Those lessons which have failed apparently need more attention from the instructional designers, but at the same time their tests need to be reviewed too because we don't know the cause of misclassifications in a test. The investigation along this line will be taken in the next section.

It should be noted that the dotted curve in Figure 2 decreases

Table 6

Estimated Number of Misclassified Students

Test	Cutoff[a]		F_+	F_-	Test	Cutoff[a]		F_+	F_-
mve103	c_0	6	5.3	1.4	mve207	c_0	5	8.5	17.2
	c	7	2.7	5.4		c	4	25.3	4.8
mve104a	c_0	7	0.2	0.0	mve301	c_0	8	6.9	4.2
	c	10	0.1	0.5		c	8	6.9	4.2
mve104b	c_0	9	2.9	2.1	mve303	c_0	15	7.6	1.9
	c	9	2.9	2.1		c	16	5.2	4.6
mve105	c_0	6	2.0	0.8	mve304	c_0	8	3.2	2.0
	c	7	1.0	3.4		c	8	3.2	2.0
mve201a	c_0	7	2.7	0.5	mve305	c_0	5	4.5	0.3
	c	8	1.8	2.0		c	7	1.8	5.7
mve201b	c_0	7	9.3	1.3	mve307	c_0	11	3.3	0.4
	c	8	6.1	4.8		c	12	2.6	1.2
mve202a	c_0	16	11.3	6.1	mve308	c_0	7	6.0	1.2
	c	16	11.3	6.1		c	8	4.1	4.7
mve202b	c_0	5	0.5	0.0	mve401	c_0	7	2.3	0.7
	c	8	0.3	1.0		c	8	1.6	2.8
mve204	c_0	8	8.8	4.4	mve402	c_0	13	14.0	3.7
	c	8	8.8	4.4		c	14	8.6	9.0
mve205a	c_0	8	12.9	12.1	mve403	c_0	8	6.5	2.5
	c	8	12.9	12.1		c	8	6.5	2.5
mve205b	c_0	8	12.4	5.2	mve404	c_0	3	14.1	0.9
	c	8	12.4	5.2		c	4	9.8	5.6
mve206a	c_0	10	3.7	1.4	mve405a	c_0	6	4.0	0.2
	c	11	2.1	4.2		c	8	2.3	3.6
mve206b	c_0	12	4.8	0.9	mve405b	c_0	8	6.9	2.9
	c	14	2.4	7.3		c	8	6.9	2.9
mve206c	c_0	13	0.4	0.0	mve405c	c_0	7	5.5	0.8
	c	16	0.2	0.8		c	8	3.6	3.3

[a] c_0 is the theoretically derived cutoff to minimize $P(F_+) + P(F_-)$.

c is the cutoff actually used in the PLATO Service Program at Chanute.

slowly for the smaller K values (No. of items in a test) but starts dropping rapidly until K reaches K=9 and again slows down. The shape of the curves varies a quite bit among MVEs and some start dropping rapidly at around K=7 or 8 for 80% true mastery level. Thus, the loss ratios of 8 and 20 can have the same optimal cutoff for the same true mastery level. This is due to that the beta binomial model deals with continuous scores while the real data are discrete.

VALIDATION OF LESSONS AND CRITERION REFERENCED TESTS

4.1 Predicting the Percentage of Success Rate for the Lesson

Table 7 shows the estimated probability of success in terms of the proportions of true score to the number of test items, or true ability level θ. These calculations are based on error free true ability level θ, so it is more reliable compared to the values obtained in Table 2., where values were calculated from the observed scores.

Since $P(\pi > .9)$, the probability of 90% of the examinees achieving mastery, was based on the observed success rate and sample size, their values don't reflect the information from tests, such as test length, 21, mean and standard deviation of a test. However, the probability P(F+ or A) is derived from unique information obtained from each test; hence we can consider it more accurate than $P(\pi \geq .9)$. The lessons which have values larger than .90 for P(A or F_-) and P(A or F_+) might not require any further revision but others might need it. Lessons 105 and 308 probably won't require any further revision, but 204,207,303,304,402, and 405b might need revision of lesson or tests in spite of not being recommended according to the validation criterion that has been used in Chanute program. The probability of PASS based on the observed scores tends to provide larger values, so that the validation criterion based on the probability of true ability level P(A or F $-$)(i.e. $p(\theta \geq \theta_0)$) will be more plausible standards.

It is important to note that these lessons may not really need revision; instead, the result may be due to poor test construction. So

Table 7

Validation Criteria

Lesson	$P(A + F_-)$	$P(A + F_+)$	$P(\pi > .90)$	Comment from Pullman et al.
103	.88	.92	.36	Recommended Monitoring (R.M.)
104a	.90	.90	.87	
104b	.86**	.87**	.08	R.M.
105	.96	.97	.25	R.M.
201a	.95	.98	.43	
201b	.94**	.97	0	Revision Recommended (R.R.)
202a	.60**	.65 **	.10	R.M.
202b	.99	.99	.97	
204	.73**	.78 **	.63	
205a	.35**	.36 **	.08	R.R
205b	.60**	.69 **	.19	R.M.
206a	.89**	.92	.03	
206b	.92	.97	.05	
206c	.99	.99	.44	
207	.41**	.31 **	.47	
301	.78**	.82 **	0	R.M.
303	.81**	.89 **		
304	.87**	.89 **	.04	
305	.92	.98	.96	
307	.93	.98	0	
308	.90	.96	0	R.R
401	.95	.98	.02	
402	.67**	.78 **	.01	
403	.66**	.70 **	.01	R.M
404	.76**	.96	.84	
405a	.94	.99	.47	
405b	.68**	.73 **	.41	
405c	.89**	.95	.63	

** recommended revision of lesson or test far, the only available technique to measure the quality of lessons is to examine the result of a CRT given at the end of the lesson. If the test is constructed very poorly (e.g. MVE 207, with $P(F_+$ or $F_-) = .2992$, $\alpha_{21}=.3287$), then the measure will be unfair to question the quality of the lesson. The measure does not distinguish between the test and the lesson. Thus, the faulty part may be the test and/or any other part or parts of the lesson. This argument can also be applied to the reverse situation. Therefore, construction of a good test will be a key point in judging the quality of a lesson that will be indirectly measured by this test.

4.2 Validation of Mastery Validation Exams

In the previous chapter, we discussed the optimal cutoff c_0 of a CRT with respect to Mastery Validation Exams in the PLATO AFB CBE Program at Chanute Air Force Base.

The evaluation study of the program, supported by Advanced Research Program Agency, measured some criterion variables which would be helpful in conducting a validation study of MVEs. The evaluation study revealed that a substantial number of examinees were miclassified(Table 6). Since detailed information on the design used in the evaluation study can be found in Dallman et al. (1977), just a brief description will be given here.

A 50-item NRT was given at the beginning and end of the eight-week PLATO AFB CBE Program, which included 37 on-line lessons. The 37 lessons were divided into four subsets called Block1, Block2, Block3, and Block4.

After a student studied and mastered all lessons in a block, he took the block test; the block test score was counted in his final grade for the course. He had to take all four block tests, and then a posttest was given in order to measure the effectiveness of the program. Each block test had twenty items which were either multiple choice or matching. The coeeficient alpha reliabilities were not calculated because the tests were writtten on the PLATO system and the item information was not collected. But α_{21} was available in the following chart. Figure 4 gives a flow chart of the testing program.

In order to validate the effectiveness of lessons, four kinds of correlations were calculated. These correlations are described in the following paragraphs.

Each Block's test scores were matched with the corresponding Master Validation Exam scores and the time needed to master the lesson (mastery time), and their correlations were calculated over the subjects. These two correlation values of 27 lessons were denoted by r(B,MVEs) and r(B,time) respectively. Their values are shown in Table 8.

The true gain scores of posttest, x_2, from pretest, x, were estimated by multiple regression procedure; the true score difference t_2-t_1 of the observed score difference x_2-x_1 was regressed on the post- and pretest scores. It is known that the regression of t_2-t_1 onto the two variables x_1 and x_2 are the same as regressing t_2-t_1 on the scores x_2-x_1 and the residual score, c_2, of x_2 on x_2-x_1 (Tatsuoka, 1975), because the covariance of x_2-x_1 and c_2 equals zero and both x_2-x_1 and c_2 are linear combinations of x_1 and x_2.

Therefore, the multiple regression $R(t_2-t_1|x_2-x_1)$ will be given

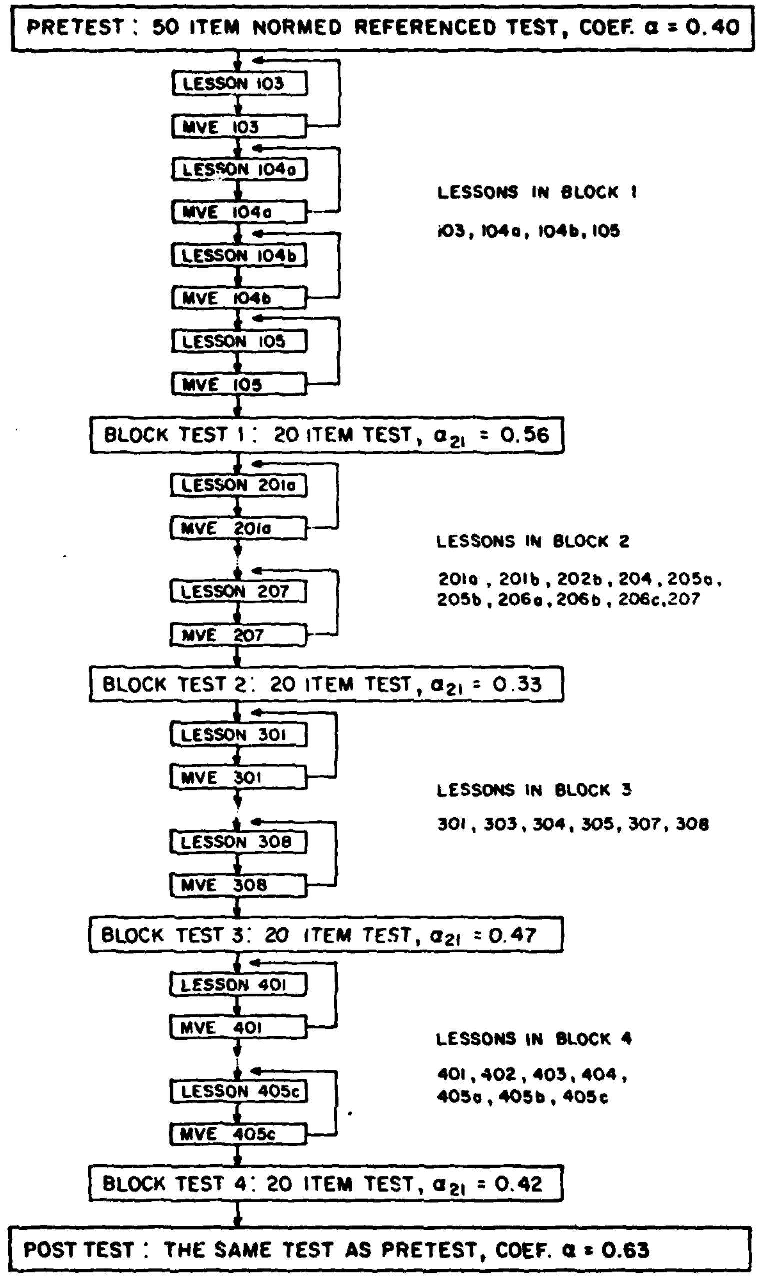


Figure 4

Block diagram of student flow through PLATO-based portion of Automotive Course as the sum of the regression of $R(t_2-t_1|\ x_2-x_1)$ and $R(t_2-t_1|c_2)$.

$$R(t_2-t_1|x_2,\ x_1) = R(t_2-t_1|x_2-x_1) + R(t_2-t_1|c_2).$$

Note that the regression coefficient of the first term is the reliability of gain scores and that of the second term is the increment of multiple R^2. The multiple R is .861, hence the reliability of the multiple regression gain score is $R^2 = .7405$. The first term, the simple difference score has the reliability of .1047 , the second term is .6358.

This estimated gain score has a higher reliability than those of pretest and posttest separately. This score was correlated with MVE scores and mastery time. Table 8 shows the result.

The optimal cutoffs that were evaluated in the previous chapter were divided by number of items in the corresponding Master Validation Exam. The same operation was used for the difference of the mean from the observed cutoff c_o in each MVE. This value expresses the distance of c_o from the mean in each test. The summary description of these variables and the correlation matrix are given in Table 9.

The probability of false positive (or advancement), P(F+) has correlation values of -.562, -.659, .638 with 'nafter', $(\text{mean}-c_o)/n$, and P(F-) (false negative or attainment), respectively. This means that the misclassification of false advancement tends to occur more often when the observed cutoff c_o is closer to the mean. The test which advances the students to the next lesson more frequently by mistake tends to retain the students whose true scores are really above the

Table 8

The correlations of Block tests to MVE scores and mastery time

lesson	r(B, MVEs)	r(B, time)	r(G, MVEs))	r(G, time)
103	.15	-.22	.23	-.38*
104a	.38*	-.33*	.19	-.43*
104b	.36*	.....	.44*	.....
105	.22	-.08	.20	-.34*
201a	.34*	.12	.44*	-.05
201b	.19	-.25	.38*	-.40*
202a	.17	-.04	.07	-.43*
202b	.26	-.03	.28*	-.07
204	.21	-.21	.11	-.13
205a	.28*	-.24	.18	-.32*
205b	.25	-.08	.15	-.26
206a	.40*	-.21	.13	-.22
206b	.12	-.04	-.02	-.18
206c	.00	-.04	.33*	-.08
207	.28*	-.17	.25	-.27
301	.04	-.08	-.11	-.06
303	.34	-.21	.08	-.05
304	.38	-.27	.42*	-.37
305	.07	-.19	.31*	-.26
307	.30*	-.23	.41*	-.30*
308	.01	.04	.00	-.07
401	.50*	-.15	.32*	-.21
402	.25	-.14	.46*	-.34*
403	.40*	-.23	.21	-.02
404	-.02	.00	.02	-.33*
405a	.07	.01	.12	-.11
405b	.25	-.06	.17	-.12
405c	.37*	-.11	.19	-.07

*significant at $p < .05$.

Table 9

A Correlation Matrix with Summary Description of Variables

	Variable	Description
1	$P(F_+)$	false positive
2	c_o/n	theoretical cutoff divided by number of items
3	$\alpha 21$	the ratio of true variance to observed variance
4	$P(F_+) + P(F_-)$	probability of misclassification
5	nafter	number of subjects using a lesson after it was declared to be validated
6	%fail	observed percentage of failure in MVE
7	$P(\pi > .9)$	Baysian estimate of success rate in the population
8	range	maximum mastery time minus minimum mastery time
9	r(G, MVEs)	correlation of gain to MVE scores
10	r(G, time)	correlation of mastery time to gain
11	r(B, MVEs)	correlation of blocktest to MVE scores
12	r(B, Time)	correlation of blocktest to mastery time
13	items	number of items in a test
14	$\frac{\text{mean} - c_o}{n}$	relative distance of c_o from the mean, c_o:observed
15	$P(F_-)$	false negative

	1	2	3	4	5	6	7	8	9	10	11
1	1.000										
2	.250	1.000									
3	-.006	.358	1.000								
4	.931	.393	-.020	1.000							
5	-.562	-.373	.037	-.617	1.000						
6	.111	.167	.384	.165	.335	1.000					
7	-.211	-.156	-.347	-.226	-.265	-.903	1.000				
8	.265	.621	.213	.345	-.304	.206	-.113	1.000			
9	-.283	-.244	.090	-.264	.271	.032	.053	-.074	1.000		
10	.183	-.233	-.259	.054	-.099	-.460	.386	-.414	-.377	1.000	
11	-.199	.051	.324	-.102	.125	.286	-.368	-.192	.403	-.275	1.000
12	.027	.053	-.316	-.056	-.133	-.320	.355	-.120	-.235	.520	-.468
13	-.108	-.271	.079	-.211	.426	.385	-.339	.070	.231	-.190	-.034
14	.659	.510	.244	.855	-.489	.408	-.396	.415	-.119	-.193	.141
15	.638	.542	.079	.869	-.544	.293	-.281	.417	-.196	-.171	.099

Note. All correlation values were transformed by Fisher's Z transformation. Probabilities were transformed by $\sin^{-1}(\sqrt{P})$.

(Table 11 cont.)

	12	13	14	15
12	1.000			
13	-.159	1.000		
14	-.228	-.119	1.000	
15	-.176	-.264	.956	1.000

mastery level. The correlation of -.659 with the variable, the number of students who studied a lesson after the validation date was set, (If over 90% of students pass the mastery level of a MVE, then the lesson was said to be validated.) indicates that the probability P(F+) will be small if the lessons whose validation date were established at an earlier date during the period of evaluation study at PLATO program.

This relation is true for the variables P(F_+ or F_-) and P(F-) because the correlations of variable nafter with them are -.617 and -.544 respectively. Moreover, P(F_+),P(F_-) and P(F_+ or F_-) correlate highly with variable (mean-c_0)/n with the values of -.659, -.855, and -.956 respectively. But the correlations between 'nafter' and (mean-c_0)/n is significant, at -.489. Hence, we cannot state that lessons which were quickly validated will produce less chance of misclassification. Since the correlation of (mean-c_0)/n and nafter is -.489, which is significantly high, the cutoff c_0 associated with some of these Mastery Validation Exams might have happened to be chosen closer to the means of corresponding MVE exams respectively. This fact raises a question about the properness of the validation criterion that has been used in PLATO Service Program at Chanute.

A stepwise multiple regression procedure was performed on the fifteen variables, and three predictors were selected to predict the variable P(F_+or F_-). Table 10 gives a summary of the analysis.

Table 10

Estimation of $P(F_+) + P(F_-)$ by Stepwise Multiple Regression

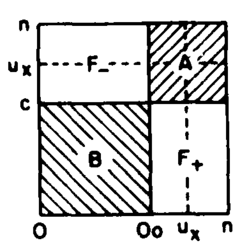
variable	coefficient	S.D. error	t
α_{21}	-.193	.088	2.193 *
nafter	-.205	.098	2.092 *
r(G, time)	.144	.089	1.618
(mean-c_o)/n	.829	.102	8.127 **

Multiple R = .9101, constant = .60, $F_{3,23} = 30.305$**

*significant at p<.05 **at p<.01

The first predictor (mean-c_o)/n for the criterion $P(F_+$ or $F_-)$, variable 4 has a beta coefficient of 0.792 and significance test of t-value 7.9. This result is expected, but entering α_{21} as the second predictor in the analysis is surprising. If α_{21} is high enough, then the probability of $P(F_+$ or $F_-)$, occurrence of misclassification, will be minimized. Most Master Validation Exams have reliabilities of around .4 to .5 which is quite low, so it is natural to expect that misclassifications will have occurred quite frequently in the program.

The variable α_{21} does not correlate significantly with variable 13, number of items in the tests; it correlates with variable 6, percentage of failure at the 5% significance level. This relationship may be interesting to investigate further , especially when the test lengths are short and about the same containing 10 - 15 items as is customary in criterion-referenced tests.



The following picture might help for quick, intuitive grasp of the relationship between F+, F- , variables c o, n and u_x. The areas of marked F+ and F- depend on on $u_x - c_o$, $n - u_x$.

Table 11

Relationship between the optimal cutoff c_0/n and other variables

variable	coefficient	S.D. error	t
α_{21}	.296	.142	2.085 *
range	.583	.141	4.135 **
no. of items	-.362	.139	2.604 *

Multiple R = .7528 , constant = .56, $F_{3,23} = 10.027^{**}$

*significant at $p<.05$ **at $p<.01$

Table 11 gives the results of a stepwise multiple regression analysis where the criterion is the optimal cutoff c_0 divided by n. Entered predictors are variables 8, 13, and 3. t-tests of significance for the beta coefficients indicate that all three variables are significant at $p<.05$. Since variable 8 is the range of time(the difference between the maximum time needed and the minimum time), the longer the time span needed by students to master a lesson, the larger the ratio of the optimum cutoff to the number of items will be. It should be noted that the procedure of evaluating the optimal cutoff c does not depend on the time needed to complete or master a lesson. But, if c/n is relatively higher, then there is more failure, both F- and correct failure, B in Figure 5, resulting a larger range in the mastery time of a lesson. α_{21} is again among the predictors and if α_{21} is larger, then c/n becomes more affected by it. This analysis needs to be more refined since a better way to interpret the results should be found.

Table 12

Relationship between r(G, MVEs) and other variables

variable	beta coefficient	S.D. error	t
c/n	-.336	.181	1.856 x
$p(\pi > .9)$	.207	.190	1.089
r(G, MVEs)	-.535	.193	2.772 *

Multiple R = .5430 , constant = 0.27, $F_{3,23}$= 3.206 *

*significant at p<.05 x significant at p<.10

Table 12 shows the results of a similar analysis, using the correlation of gain scores and Mastery Validation Exam scores as the criterion. A larger value of this variable means that the gain score was non-negligibly affected by the Mastery Validaion Exams, which have a large correlation value r(G,MVEs). We know from Table 10 that MVE scores of lessons 104b, 201a, 201b, 206c, 304, 305, 307, 401, and 402 have significant values of correlation. This analysis revealed that correlation of mastery time to gain scores contributes the most significantly in predicting variable 9. Since mastery time of a lesson correlates highly with aptitude scores as shown in Table A of the Apppendix, this result is expected.

The students affected most by the decision of cutoff scores are mediocre students whose scores are near the cutoffs, and therefore they tend to be more often misclassified in either the positive or negative way. The fact that the beta coefficient of variable 2 is -.336 means that the smaller the values of c_0/n, the larger the contribution to the gain will be; thus mediocre students have a greater chance of repeating the lessons since the observed cutoff c was set to 80% across all MVEs, which is the true mastery level that was aimed for.

Table 13

Relationship Between $p(\pi > .9)$ and other variables.

variable	beta coefficient	S.D. error	t
21	-.152	.178	.854
r(G, MVEs)	.224	.185	1.211
r(G, time)	.305	.190	1.605
no. of items	-.344	.195	1.966 x
$(\text{mean}-c_o)/n$	.314	.199	1.954 x

Multiple R = .6503 , constant = 1.09, $F_{5,21} = 3.077$ *

*significant at p<.05 x significant at p<.10

Table 13 shows the results of analysis when the criterion is variable 9, probability $P(\pi \geq .9)$ that 90% or more of the students in the

*next page

population from which our sample was drawn will achieve the 80% mastery level on the end of lesson test. Five predictors among variables 1, 2, 3, 4, 8, 9, 10, 11, 12, 13, 14, and 15 were selected. The variables nafter and % fail were omitted because $P(\pi > .9)$ was derived from these two values in the sample. None of the beta coefficients was significant, but we might be able to say that $P(\pi > .9)$ depends to some extent on the test length (beta=-.344, t=1.97). Also, the distance of the mean from the observed cutoff c_o affects the value of $p(\pi > .9)$ such that if the observed cutoff c_o is considerably smaller than the mean, then the success rate of the lesson becomes larger. This means that the test was probably too easy in comparison with other tests. This analysis result confirms that the validation criterion used at the PLATO AFB CBE program at Chanute Air Force Base depended excessively on the test, the characteristics of MVE; hence the method that was used to assess the quality of lessons was inadequate. There is a great need for the development of a method to validate lessons directly, without depending entirely on the end of lesson tests.

SUMMARY AND DISCUSSION

The problem of setting a validation criterion for a given lesson is important in practice, but it has never become a focus for educational researchers, although the closely related topic of criterion referenced test has been one of the most popular research targets in the past few years. Both the sample binomial model and the Bayesian binomial model (beta binomial model) are adopted to set a better validation criterion for a given lesson and the result from the latter model matched our data better than did the former. Therefore, the prediction of the future success rate of the lesson using the Bayesian binomial model is recommended for setting a validation criterion, when (a) the information is limited to the percentage of failure (or success) rate on the end of the lesson test and (b) an author (or instructor) of the lesson has a certain level of prior belief as to what extent his/her lesson will be successful. If the scores of a test given at the end of a lesson are available, then it is recommended to use the information that one can get from the test performance as much as possible upon setting a validation criterion of the lesson. Applying the beta binimial model of criterion-referenced testing, the estimated probability of the observed score X being larger than the observed cutoff c will be a better validation criterion than the success rate. In other words, the probability of mastery, passing the criterion score c will serve as a validation criterion of the lesson.

Of course, the decision of mastery or non-mastery must theoretically be based on a student's true performance level and not on the observed scores, but the true score will never be available in practice. But it is possible to estimate the probability of the true score being greater than or equal to a given true mastery level, say, 80%. Unfortunately, we don't have any analytical method to determine the best, most suitable true mastery level for a program.

The four kinds of probabilities -- correct pass (A), correct fail (B), false positive (F+) and false negative (F-) -- were calculated over 27 Mastery Validation Examinations (a) when the observed cutoff c, (80% correct) and (b) when the optimum cutoff c_o, which minimizes misclassification of students, was used. The results indicate that even if c_o were used in the decision process, some tests still show substantially large numbers of misclassifications of both the false positive and false negative types. Since it is interesting to investigate why some tests showed as much as about 20 % of misclassification while other tests showed very little, three stepwise multiple regression analyses were used to select the predictors of P(F+), P(F-), and P(F+ or F-) separately. The common strongest predictor was the distance of c_o from the mean of a test, which was what we expected. The second common predictor was $_{21}$, the internal consistency of a criterion referenced test. As $_{21}$ increases to 1, all three criterion variables get smaller, hence less misclassifications occur. That means the internal consistency of the items in a given test is important to control false positive and false negative errors.

The optimum cutoff c_o's for Mastery Validation Exams are smaller than or equal to the actually used observed cutoff c's in almost all cases in the PLATO AFB CBE project. Therefore the probabilities of false negative associated with c_o are smaller than or equal to those of false negative associated with the observed cutoff c. But the probabilities of false positive associated with c_o tend to be larger than those associated with c. Since we set the loss ratio to 1 in this case, the total probability of misclassification is always minimized by using the optimum cutoff c_o. P(F+) in some test is eight times as large as P(F-), while in others the former is only a few times larger. Setting the most appropriate loss ratio will be a problem when Huynh's method to evaluate the optimum cutoff is adopted. Also, his method is more sensitive for the smaller loss ratios than larger ones, say Q=10-20. Our data showed that many Master Validation Examinations of the end-of-lesson tests had the same optimal cutoff c_o for loss ratios between 8 and 20. If his intention was to control the false positive errors upon the decision of mastery-non mastery for a linearly related curriculum such as mathematics, then the applicability of the method in educational settings will be a problem.

APPENDIX

Table A

Correlations of Aptitude Scores with MVE Scores,

First Completion Time, Mastery Time, and Test Completion Time

Lesson	MVE scores	First completion time	Mastery time	Test completion time
103	.45*	-.39*	-.08*	-.32*
104a	.17	-.40*	-.38*	-.06
104b	time data was lost			
105	.31*	-.42*	-.49*	-.32*
201a	.52*	.04	-.08	-.32*
201b	.16	-.42*	-.42*	-.33*
202a	.38*	-.12	-.25	-.10
202b	.34*	-.19	-.19	-.42*
204	.19	-.16	-.22	-.26
205a	.39*	-.38*	-.45*	-.32*
205b	.47*	-.00	-.27	-.20
206a	.42*	-.03	-.14	-.42*
206b	.27	-.25	-.27	-.22
206c	.24	.02	-.02	-.40*
207	.24	-.23	-.26	-.15
301	.24	-.03	-.13	-.34*
303	.10	-.39*	-.26	-.19
304	.60*	-.14	-.36*	-.51*
305	.17	-.35*	-.36*	-.45*
307	.52*	-.54*	-.59*	-.57*
308	.20	-.00	-.03	-.54*
401	.38*	-.41*	-.41*	-.39*
402	.47*	-.27	-.39*	-.39*
403	.48*	-.24	-.31*	.09
404	.10	-.27	-.27	-.32*
405a	.27	-.15	-.27	.12
405b	.05	-.03	-.19	-.05
405c	.31*	-.11	-.06	.02

*$p < .05$

TABLE B

Description of Contents in the Lessons of Chanute

lesson	Content
103	Principles of Gas Engine
104a, 104b	Identification of Parts and Purpose of Gasoline Engine Compressor
105	Cooling System
201a, 201b	Air and Exhaust System
202a	Fundamentals of Electricity
202b	Batteries
203a, 203b, 203c	Electrical Schematics
205a	Cranking Motors, DC Charging System
205b	AC Charging System
206a, 206b, 206c	Battery Ignition
207	Emission Control
301	Diesel Engines
303	Lighting System
304	Warning System
305	Clutches
307	Basic Hydraulics
308	Fluid Couplings/Torque Converters
401	V-Joints/Propeller Shafts
402	Differentials
403	Transfer Case/PTO
404	Suspension System
405a	Hydraulic and Mechanical Brakes
405b	Air Brakes
405c	Power Assisted Brakes

REFERENCES

Atkinson, R.C., Computer-based instruction in initial reading. In proceedings of the 1967 invitational conference on testing problems. Princeton, Educational Testing Service, 1968, 58-67.

Besel, R., A comparison of Emrick and Adam's mastery-learning test model with Kriewall's criterion-referenced test model. Inglewood, California: Southwest Regional Laboratory, Technical Memorandum 5-71-04, April, 1971.

Block, J. H., The effects of various levels of performance on selected cognitive, affective, and time variables. Unpublished Ph.D. dissertation, University of Chicago, 1970.

Block, J.H., (Ed.) Mastery learning: theory and practice. New York: Holt, Reinhart & Winston, 1971.

Block, J.H., Student learning and the setting of mastery performance standards. Educational Horizons, 1972, 50, 183-190

Bloom, B.S., Learning for mastery, UCLA-CSEIP Evaluation Comment, 1, 2, 1968

Branson, R.K., et al. Interservice procedures for instruc tional system development: Phase III. Florida State University, August, 1975.

Carrol, J.B., A model of school learning. Teachers College Records, 1963, 64, 723-733.

Carroll, J. B., & Spearitt, D., A study of a model of school learning. Monograph No.4 Cambridge, Massachusetts: Harvard University, Center for Research and Development of Educational Differences, 1967.

Dallman, B.E., Deleo, P.J., Main, P.S., and Gillman, D.C., Evaluation of PLATO IV in vehicle maintenance training. AFHRL-TR-77-59, Lowry AFB CO; Technical Training Division, Air Force Human Resources laboratory, November, 1977.

Emrick, J. A. An Evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Ferguson T.S., Mathematical statistics: A decision theoretic approach. New York: Academic Press, 1967.

Glaser, R., Instructional technology and the measurement of learning outcomes--some questions, American Psychologist, 1963, vol. 18, 519-521.

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 43-64.

Keats, J.A. & Lord, F.M., A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72

Kim, Hogwon, et al. The mastery learning project in the middle schools. Seoul: Korean Institute for Research in the Behavioral Science, 1970.

Linn, R.L. Personal communication, October 1, 1977.

Lord, F.M. & Novick, M.R., Statistical theories of mental test scores. Reading: Adison-Wesley,1968.

Millman, J. Tables for determining number of items needed on domain-referenced tests and number of students to be tested. Los Angeles: Instructional Objectives Exchange, Technical Paper No.5, April 1972.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973,43, 205-215.

Novick, M.R. & Jackson, P.H., Statistical methods for Educational and psychological research. New York: McGraw-Hill, 1974.

Novick, M. R. & Lewis, C. Prescribing test length for criterion-referenced measurement. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angels: UCLA Graduate school of Education, Center for the Study of Evaluation, 1974.

Roudabush, G. E., Item selection for criterion-referenced tests. Paper presented at Annual meeting of American Educational Researdh Association, New Orleans, 1973.